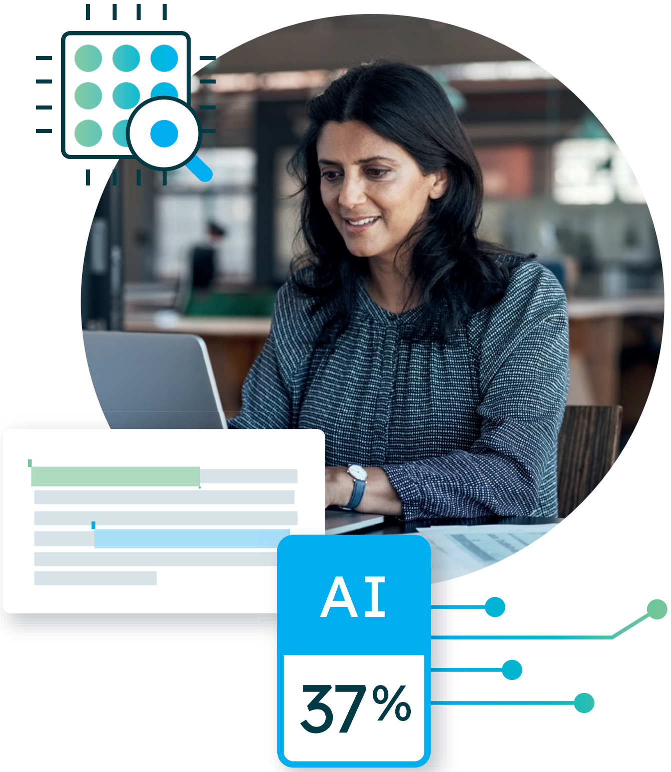
FAQs for Turnitin’s AI writing detection capabilities



How do Turnitin's AI writing detection capabilities work?

1. Does Turnitin offer a solution to detect AI writing?

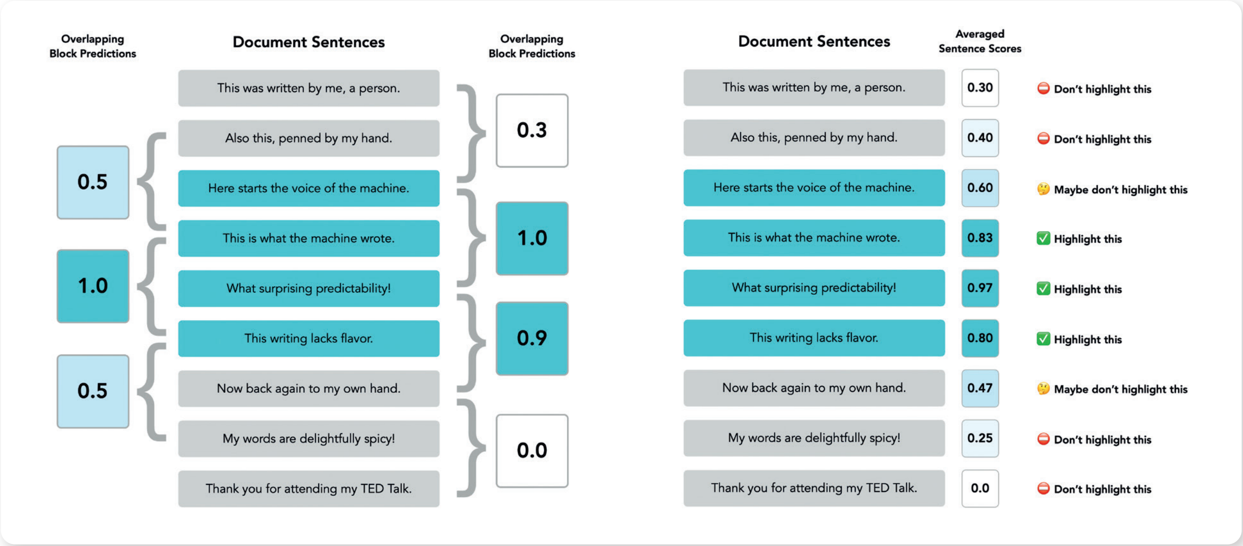
Yes. Turnitin has released its AI writing detection capabilities to help educators uphold academic integrity while ensuring that students are treated fairly.

We have added an AI writing indicator to the Similarity Report. It shows an overall percentage of the document that AI writing tools, such as ChatGPT, may have generated. The indicator further links to a report which highlights the text segments that our model predicts were written by AI. Please note, only instructors and administrators are able to see the indicator.

While Turnitin has confidence in its model, Turnitin does not make a determination of misconduct, rather it provides data for the educators to make an informed decision based on their academic and institutional policies. Hence, we must emphasize that the percentage on the AI writing indicator should not be used as the sole basis for action or a definitive grading measure by instructors.

2. How does the solution work?

When a paper is submitted to Turnitin, the submission is first broken into segments of text that are roughly a few hundred words (about five to ten sentences). Those segments are then overlapped with each other to capture each sentence in context.


Overlapping Block Predictions
Document Sentences
Overlapping Block Predictions
0.5
1.0
0.5
This was written by me, a person.
Also this, penned by my hand.
Here starts the voice of the machine.
This is what the machine wrote.
What surprising predictability!
This writing lacks flavor.
Now back again to my own hand.
My words are delightfully spicy!
Thank you for attending my TED Talk.
0.3
1.0
0.9
0.0
Document Sentences
Averaged Sentence Scores
This was written by me, a person. 0.30 ⛔ Don't highlight this
Also this, penned by my hand. 0.40 ⛔ Don't highlight this
Here starts the voice of the machine. 0.60 🤔 Maybe don't highlight this
This is what the machine wrote. 0.83 ✅ Highlight this
What surprising predictability! 0.97 ✅ Highlight this
This writing lacks flavor. 0.80 ✅ Highlight this
Now back again to my own hand. 0.47 🤔 Maybe don't highlight this
My words are delightfully spicy! 0.25 ⛔ Don't highlight this
Thank you for attending my TED Talk. 0.0 ⛔ Don't highlight this

The segments are run against our AI detection model and we give each sentence a score between 0 and 1 to determine whether it is written by a human or by AI. If our model determines that a sentence was not generated by AI, it will receive a score of 0. If it determines the entirety of the sentence was generated by AI it will receive a score of 1.

Using the average scores of all the segments within the document, the model then generates an overall prediction of how much text (with 98% confidence based on data that was collected and verified in our AI Innovation lab) in the submission we believe has been generated by AI. For example, when we say that 40% of the overall text has been AI-generated, we're 98% confident that is the case.

Currently, Turnitin's AI writing detection model is trained to detect content from the GPT-3 and GPT-3.5 language models, which includes ChatGPT. We are actively working on expanding our model to enable us to better detect content from other AI language models.

3. What parameters or flags does Turnitin's model take into account when detecting AI writing?

GPT-3 and ChatGPT are trained on the text of the entire internet, and they are essentially taking that large amount of text and generating sequences of words based on picking the next highly probable words. This means that GPT-3 and ChatGPT tend to generate the next word in a sequence of words in a consistent and highly probable fashion. Human writing, on the other hand, tends to be inconsistent and idiosyncratic, resulting in a low probability of picking the next word the human will use in the sequence.

Our classifiers are trained to detect these differences in word probability and are adept to the particular word probability sequences of human writers.

4. How was Turnitin's model trained?

Our model is trained on a representative sample of data that includes both AI-generated and authentic academic writing. While creating our sample dataset, we took into account statistically under-represented groups like second-language learners, English users from non-English speaking countries, students at colleges and universities with diverse enrollments, and less common subject areas such as anthropology, geology, sociology, and others.

5. Can I check past submitted assignments for AI writing?

Yes. Previously submitted assignments can be checked for AI writing detection if they're re-submitted to Turnitin. Only assignments that are submitted after the launch of our capability (4th April 2023) are automatically checked for AI writing detection.

6. What languages are supported?

English. For the first iteration of Turnitin's AI writing detection capabilities, we are able to detect AI writing for documents submitted in long-form English only.

7. What will happen if a non-English paper is submitted?

If a non-English paper is submitted, the detector will not process the submission. The indicator will show an empty/error state with 'in-app' guidance that will tell users that this capability only works for English submissions at this time. No report will be generated if the submitted content is not in English.

8. Can my institution get early access to be able to trial this new capability?

No. Unlike the usual Turnitin product launches, we're unable to provide early access for trial purposes to any customers for this release as we're bringing this technology to market at an accelerated speed, based on customer feedback. However, we have been rigorously testing our AI detection technology in our labs and are confident of the results.

In order to help customers understand this capability, we will be providing documentation and step-by-step guidance on how to use it.

9. Can I or my admin suppress the new indicator and report if we do not want to see it?

No. For this first iteration, we're unable to suppress the AI writing detection indicator or the report for any customer using Turnitin Feedback Studio (TFS), TFS with Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check, and Originality Check+.

We have received a significant amount of positive feedback from customers to move quickly with our AI writing detection technology in order to create visibility when it comes to students using AI-writing tools and provide insights to educators as soon as possible. Adding suppression capability would increase our time to market.

We understand that your role as administrators is to support instructors, and we recognize your concern that some instructors may misinterpret the results. To address this, we have taken measures to reduce the support burden on you for this new capability and to help you educate instructors on how to interpret the percentage indicated by our AI writing indicator. These measures include:

- In-product guidance to help institutions understand our AI detection capabilities.
- Link to an FAQ page from within the AI writing detection report, explaining what the various results mean and how they should be interpreted.
- **Slide deck** explaining how the feature works, to share with others at your institution.
- **Additional FAQs** that explain more about the feature, results, licensing, etc.
- Access to a number of **pedagogical resources** to help educators understand how to uphold academic integrity in the age of AI.

10. Will the addition of Turnitin's AI detection functionality to the Similarity Report change my workflow or the way I use the Similarity Report?

No. This additional functionality does not change the way you use the Similarity Report or your existing workflows. Our AI detection capabilities have been added to the Similarity Report to provide a seamless experience for our customers.

11. Will the AI detection capabilities be available via LMSs such as Moodle, Blackboard, Canvas, etc? What about Microsoft Teams?

Yes, users will be able to see the indicator and the report via the LMS they're using. We have made AI writing detection available via the Similarity Report. There is no AI writing indicator or score embedded directly in the LMS user interface and users will need to go into the report to see the AI score.

Please note however, that AI writing detection will not be available via the Microsoft Teams integration. This is because the current MS Teams integration only uses the student viewer. There is no separate teacher viewer. And since our AI detection capability is only going to be available to educators, we cannot provide it via MS Teams.

12. How is authorship detection within Originality different from AI writing detection?

Turnitin's AI writing detection technology is different from the technology used within Authorship (Originality). Our AI writing detection model calculates the overall percentage of text in the submitted document that was likely generated by an AI writing tool. Authorship, on the other hand, uses metadata as well as forensic language analysis to detect if the submitted assignment was written by someone other than the student. It will not be able to indicate if it was AI written; only that the content is not the student's own work.

AI detection results & interpretation

1. What does the percentage in the AI writing detection indicator mean?

The percentage indicates the amount of qualifying text within the submission that Turnitin's AI writing detection model determines was generated by AI (with 98% confidence based on data that was carefully collected and verified in a controlled lab environment). This qualifying text includes only prose sentences, meaning that we only analyze blocks of text that are written in standard grammatical sentences and do not include other types of writing such as lists, bullet points, or other non-sentence structures.

This percentage is not necessarily the percentage of the entire submission. If text within the submission is not considered long-form prose text, it will not be included.

2. The percentage shown sometimes doesn't match the amount of text highlighted. Why is that?

Unlike our Similarity Report, the AI writing percentage does not necessarily correlate to the amount of text in the submission. Turnitin's AI writing detection model only looks for prose sentences contained in long-form writing. Prose text contained in long-form writing means individual sentences that make up a longer piece of written work, such as an essay, a dissertation, or an article, etc. The model does not detect AI-generated text such as poetry, scripts, or code. Nor does it detect short-form/unconventional writing such as bullet points, tables, or short exam answers.

3. What is the accuracy of Turnitin's AI writing indicator?

We only flag something as AI-written when we are 98% sure it is written by AI. This is because we want to make sure we don't falsely flag something as AI-generated that isn't. This means, however, that we will likely miss up to 15% of text written by AI, with a less than 1% false positive rate (incorrectly identifying fully human-written text as AI-generated).

For example, if we identify that 50% of a document is written by AI, we are 98% sure that at least 50% is written by AI with a less than 1% false positive rate, but it could contain as much as 65% AI writing.

The above rates have been determined by our model using data that was collected and verified in our AI Innovation Lab, but we know that real world use will differ from the lab tests. To take this into account, we've tuned our AI detector to minimize false positives on authentic text, even if it means we might miss some instances of AI writing.

4. What can I do if I feel that the AI indicator is incorrect? How does Turnitin's indicator address false positives?

If you find AI written documents that we've missed, or notice authentic student work that we've predicted as AI-generated, please let us know! Your feedback is crucial in enabling us to improve our technology further. You can provide feedback via the 'feedback' button found in the AI writing report.

Sometimes false positives (incorrectly flagging human-written text as AI-generated), can include lists without a lot of structural variation, text that literally repeats itself, or text that has been paraphrased without developing new ideas. If our indicator shows a higher amount of AI writing in such text, we advise you to take that into consideration when looking at the percentage indicated.

In a longer document with a mix of authentic writing and AI generated text, it can be difficult to exactly determine where the AI writing begins and original writing ends, but our model should give you a reliable guide to start conversations with the submitting student.

In shorter documents where there are only a few hundred words, the prediction will be mostly "all or nothing" because we're predicting on a single segment without the opportunity to overlap. This means that some text that is a mix of AI-generated and original content could be flagged as entirely AI-generated.

Please consider these points as you are reviewing the data and following up with students or others.

5. Will students be able to see the results?

The AI writing detection indicator and report are not visible to students.

6. What is the difference between the Similarity score and the AI writing detection percentage? Are the two completely separate or do they influence each other?

The Similarity score and the AI writing detection percentage are completely independent and do not influence each other. The **Similarity score** indicates the percentage of matching-text found in the submitted document when compared to Turnitin's comprehensive collection of content for similarity checking.

The AI writing detection percentage, on the other hand, shows the overall percentage of text in a submission that Turnitin's AI writing detection model predicts was generated by AI writing tools.

7. Does the Turnitin model take into account that AI writing detection technology might be biased against particular subject-areas or second-language writers?

Yes, it does. One of the guiding principles of our company and of our AI team has been to minimize the risk of harm to students, especially those disadvantaged or disenfranchised by the history and structure of our society. Hence, while creating our sample dataset, we took into account statistically under-represented groups like second-language learners, English users from non-majority-English countries, students at historically black colleges and universities, and less common subject areas such as anthropology, geology, sociology, and others.

8. How can I use the AI indicator percentage in the classroom with students?

Turnitin's AI detection indicator shows the percentage of text that has likely been generated by an AI writing tool while the report highlights the exact segments that seem to be AI-written. The final decision on whether any misconduct has occurred rests with the reviewer/instructor. Turnitin does not make a determination of misconduct, rather it provides data for the educators to make an informed decision based on their academic and institutional policies.

Scope of detection

1. Which AI writing models can Turnitin’s technology detect?

The first iteration of Turnitin’s AI writing detection capabilities have been trained to detect models including GPT-3, GPT-3.5, and variants. Our technology can also detect other AI writing tools that are based on these models such as ChatGPT. We plan to expand our detection capabilities to other models in the future.

2. Is your current model able to detect GPT-4 generated text?

We are constantly working on improving and expanding our AI writing detection capabilities. Currently, our AI team is conducting tests on GPT-4 using our existing detector to compare its performance and understand the differences between GPT-3.5 (on which our model is trained), and GPT-4. Preliminary results are promising as we’re accurately detecting AI-generated text. Our analysis is ongoing and once we have established reliable efficacy metrics, we will update our models to include GPT-4. It is important to note, however, that the free version of ChatGPT is still operating on GPT-3.5.

3. How will Turnitin be future-proofing for advanced versions of GPT and other large language models yet to emerge?

We recognize that Large Language Models (LLMs) are rapidly expanding and evolving, and we are already hard at work building detection systems for additional LLMs. Our focus initially is on building and releasing an effective and reliable AI writing detector for GPT-3 and GPT-3.5, and other writing tools based on these models such as ChatGPT.

4. Can Turnitin detect if text generated by an AI writing tool (ChatGPT, etc.) is further paraphrased using a paraphrasing tool? Will it flag the content as AI-generated even in this instance?

Our detectors are trained on the outputs of GPT-3, GPT-3.5 and ChatGPT, and modifying text generated by these systems will have an impact on our detectors’ abilities to identify AI written text. In our AI Innovation Lab, we have conducted tests using open sourced paraphrasing tools (including different LLMs) and in most cases, our detector has retained its effectiveness and is able to identify text as AI-generated even when a paraphrasing tool has been used to change the AI output.

5. Does Turnitin have plans to build a solution to detect when students paraphrase content either themselves or through tools such as Quillbot, etc.,?

Turnitin has been working on building paraphrase detection capabilities – ability to detect when students have paraphrased content either with the help of paraphrasing tools or re-written it themselves – for some time now, and the technology is already producing the desired results in our AI Innovation Lab. In the instance when the student is using a word spinner or an online paraphrasing tool, the student is just running content through a word spinner which uses AI to intentionally subvert similarity detection, not using generative AI tools such as ChatGPT to create content.

We have plans for a beta release in 2023, and we will be making paraphrase detection available to instructors at institutions that are using TFS with Originality and Originality for an additional cost. It will be released first in our TFS with Originality product.

Access & licensing

1. Who will get access to this solution? Will we need to pay more for this capability?

The first iteration of our AI writing detection indicator and report are available to our academic writing integrity customers as part of their existing licenses, so that they're able to test the solution and see how it works. This includes customers with a license for Turnitin Feedback Studio (TFS), TFS with Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check, and Originality Check+. It is available for customers using these platforms via an integration with an LMS or with Turnitin's Core API. Please note, only instructors and administrators will be able to see the indicator and report.

Beginning January 1, 2024, only customers licensing Originality or TFS with Originality will have access to the full AI writing detection experience.

2. When can customers get access to this solution?

Turnitin's AI writing detection capabilities are available now and have been added to the Similarity Report. Customers licensing any of the above Turnitin products should be able to see the indicator and access the AI report.

3. Is Turnitin's AI writing detection a standalone solution or is it part of another product?

The first iteration of Turnitin's AI writing detection capabilities is a separate feature of the Similarity Report and is available across these products: Turnitin Feedback Studio (TFS), TFS with Originality, Turnitin Originality, Turnitin Similarity, Simcheck, Originality Check, and Originality Check+. The indicator links to a report which shows the exact segments that are predicted as AI-written within the submitted content.

4. Why is AI detection not being added to other Turnitin products like Gradescope and iThenticate?

We focused our resources on, what we view, as the biggest, most acute problem and that is higher education and K12 long-form writing. We are currently investigating how we can bring AI writing detection to iThenticate customers. We do not currently have plans to add these capabilities to Gradescope, since the primary use case for Gradescope is handwritten text while for AI detection we're focusing on typed text. However, we are happy to learn more about customer needs for AI writing detection within this product. In addition, we are not pursuing ChatGPT code detection at this time.

5. Where can I find more information about this new solution?

You can find information about Turnitin's AI writing detection capabilities **on this page**.